# 日本語名詞に対する疑問詞タグ辞書の作成

山本 和英　　後藤 大明

長岡技術科学大学

{yamamoto, goto}@jnlp.org

## １．普通名詞に関する言語資源が不足している

今年度から我々は「雪だるま」プロジェクトを開始した(山本ら 2015)。このプロジェクトでは、自然言語処理の底上げのためには従来利用可能な言語資源だけでは圧倒的に不十分であるという危機意識の下で、様々な辞書の作成、これを統合的に組み込んだ解析器の作成、及び一般公開を目指している。すでに単語解析器については試作を行って単語体系や品詞体系の見直しを進めており、また略語を含む表記ゆれの吸収(Yamamoto et al. 2015)[1]や同義語の正規化の作業(Yamamoto and Takahashi 2015) を継続して行っている。

本研究では日本語普通名詞に対して最も大まかな意味情報である「疑問詞タグ」を付与した辞書を作成した。疑問詞タグとは、「いつ」「どこ」「誰」「何」「どれだけ」のそれぞれの質問について、各名詞が回答になり得るかどうか(すなわち、疑問詞と名詞が同一の属性を有しているか)を名詞ごとに5つの二値で示した情報である。

後述するように、このような情報は名詞の最も基本的な特性を表現しており、様々なタスクにおいて性能の劣化なく正解候補の絞り込み(＝計算量の削減、性能の向上)を可能にする重要な情報である。しかしながら、このような辞書は従来存在せず[2]、またこれに相当する情報を含む言語資源も間接的、部分的にしか存在しない。そこで、我々は今回この付与を試みたのでその必要性と作業内容について報告、議論する。

## ２．疑問詞タグ

### 2.1　疑問詞タグは何に使うのか？

**[質問応答]** 事実を問うファクトイド型の質問応答では、「いつ」「どこ」「何」といった質問が入力され、これに対して通常は構造化されていない大規模テキストから回答を探索する必要がある。ここで、出現する全名詞を回答候補にしている(最近の研究として(松井ら 2015;p.176)では名詞、複合名詞、アルファベットの連続、形容詞を対象にしている)。ここで、予め回答になり得る名詞の部分集合があれば探索空間を減らすことでき、計算量の低減と誤りの減少が期待できる。

**[照応解析・主体推定]** 代名詞で表現された「彼」「彼女」などが何を指すかを解析するのが照応解析である。特に人を表す代名詞の先行詞となり得る語は限定的で、すべての名詞が対象となる訳ではない。主体推定も同様で、例えば(叶内ら 2015;p.208)では、主体として適切な単語(例:彼女、社員、部下)とその他の人物(例:幼児)の辞書をそれぞれ人手で作成して主体推定器の素性として用いている。

**[格解析などの意味解析]** 例えばニ格の深層格推定(竹野ら 2014)において時間格(「いつ」)や場所格(「どこ」)になり得る名詞の集合が予め得られていれば、精度向上が期待できる。

以上のように、疑問詞タグは名詞群を分類するための最も基本的とも言える属性であり、これ以外にも語句の意味が関係する様々なタスク、例えば語義曖昧性解消などで利用できる。

---

[1] (Yamamoto et al. 2015)以降も、旧漢字語の新漢字語への統合や機能表現の統合など、表記統制のための辞書構築を継続して行っている。

[2] 固有表現に対しては同様の情報が整備されつつあるのに、中核的な普通名詞に対してこれまで整備されてこなかったのは不思議に思える。

### 2.2 シソーラスで十分ではないか？

次に、従来研究を俯瞰することで本研究の必要性を議論する。

現状の日本語処理において事実上唯一とも言える意味的な言語資源がシソーラス（is-a オントロジー）である。シソーラスは主要な語彙を何らかの基準で階層的に分類した言語資源であり、分類の結果として木構造で表現される。日本語のシソーラスとしては日本語語彙大系、日本語 WordNet、分類語彙表、角川類語新辞典などが知られている。

シソーラスは語句の意味分類なのだから、一見すると我々が望む疑問詞タグに相当する情報もシソーラスから得られそうに感じる。例えば、「誰」タグは人間に関係ある分類に属する語の集合で良さそうである。分類語彙表を例にすると、各分類について分類名【…】とさらに下位分類の冒頭語を以下に示す。これらをすべて「誰」と呼んでいいかは若干議論の余地があるが、概ね下記分類に属する語で良さそうである[3]。

1.20【人間】人間、我、自他、神仏、男女、長幼
1.21【家族】家族、夫婦、親、子、はらから、親戚
1.22【仲間】相手、師友、主客
1.23【人物】人種、国民、元首、皇族、霊長、佐藤
1.24【成員】～員、技師、官憲、事業家、農民、航海士、職人、警察官、使用人、学徒、軍人、長、君臣、関係者

しかし、「どこ」タグになるとシソーラスの分類から作成するのは容易ではない。場所を表すのは物理的な単語だからと言って物体すべてが「どこ」になる訳でもなく、逆に通常物体とは認識されない組織名や地名に準ずる普通名詞も「どこ」の対象になり得る。

さらに重要なのは、シソーラスは原則として分類になっている（多義性がある場合を除いて1つの語が複数の分類に含まれることはない）のに対して、疑問詞タグは相互に排他的でないということである。タグ付与結果の節で述べるが、1つの語に複数のタグが付与されることは決して珍しいことではなく、また不自然でもない。タグ間の関係が排他的になっていないのであれば、シソーラスから各タグに含まれる分類項目はそれぞれのタグごとに検討する作業が発生するので、シソーラスから容易に得られるという話にはならない。

我々は以前に、Wikipedia 記事を自動分類する際に施設と組織の弁別が難しいことを報告した（柴木ら 2012;p.254）。施設は「どこ」、組織は「なに」の対象であることに注意すると、このタスクは本稿での作業に部分的に対応する。当該論文のタスクでは両者をどちらかに決めていたが、ある名詞群については本質的に両者の性質を持っていると考えたほうが自然である。これは多義性とは別の概念であり、我々はこれを多面性と呼んでいる。従来のシソーラスは名詞の一面だけを捉えて分類しており、名詞の多面性を表現できていない。これを言語資源として顕在化させようという試みの第一歩が今回の作業である。

以上まとめると、シソーラス上には疑問詞タグに関連する情報が含まれていることは間違いなく、またシソーラスを用いることで類似した語についてまとめて作業できるので効率的でもある。しかしながら、シソーラスを用いるにせよ何らかの加工作業が必要であり、少なくともシソーラスから容易に疑問詞タグを抽出できる状態にはなっていない。

## 3. 疑問詞タグの付与

### 3.1 付与作業

まず、本作業で付与する疑問詞タグは下記の5種類とした。疑問詞はこの他に「なぜ」「どのように」などがあるがこれらは対象外とした。

「いつ」（時間）
「どこ」（場所）
「だれ」（ヒト）
「なに」（モノ）
「どれだけ」（数量、規模）

作業は、名詞に対して各疑問詞の対象になるかどうかを二値(0, 1)で付与した。例えば、「明日」という単語は「いつ」の回答になり得るので「いつ」に対して1、その他のタグは 0 となる。同様に「学校」には対しては「どこ」「なに」が 1、その他は 0 となる。作業は、著者1

[3] ちなみに、これらは分類番号 1.2 のすべてではない。1.25【公私】1.26【社会】1.27【機関】は明らかに「誰」タグの対象外である。つまり分類語彙表は単独カテゴリで「誰」タグを表現できていない。

名（後藤）の主観のみで判断し、作業時にシソーラス等の参照は行っていない。

情報の付与は常識的な世界知識の下で行った。物語や夢などの空想の世界では様々な名詞を擬人化して使用することで動物や無生物も「だれ」の対象になり得るが、このような場合は考えずに情報を付与した。同様に、普通名詞と同名の地名や人名についても、一般的に知られていない限りは可能性を排除した。例えば「朝」という名詞は普通名詞としてのみ考えて付与し、人名や地名としての可能性は作業者が特に思いつかない限り排除した。

対象とする名詞は日本語の高頻度普通名詞1万5千語とした。ここで、複合名詞は普通名詞とは別の語として付与対象となっていることに注意されたい。頻度情報は、現代日本語書き言葉均衡コーパス(BCCWJ)全文に対して我々が開発している日本語解析システム「雪だるま」を用いて単語解析を行い、その結果から獲得した。以上のようにして具体的な名詞選定を行っているため、当該言語資源は完全な形で単語解析器と統合することが可能であり、従来研究のようにツールと資源の単語分割基準の違いに悩まされることが全くない。これは解析器と言語資源の統合を目指すべきであるという雪だるまプロジェクトの大きな主張の一つである。

### 3.2 付与結果

タグ付与結果を表1に示す。前述したように作業語数は1万5千語であるが、タグの重複やタグなしの語があるので表1での件数の総和は1万にならない。

## 4. 議論

### 4.1 疑問詞タグの被覆率

3.1 節で述べたように、今回は１万５千語の名詞を対象に本作業を行った。これによって、BCCWJ の全出現名詞数の約７割に対して疑問詞タグを付与することができた。我々は最終的には９割程度の被覆率が必要と考えており、現在の被覆率は改善の余地がある。

表1:疑問詞タグ付与結果

| 疑問詞 | 例 | 件数 |
|---|---|---|
| いつ | 何時、最近、 | 320 |
| どこ | 外国、家庭、ドア、都市、銀行、会場 | 1,608 |
| だれ | 我々、２人、父さん、医師、自分たち、兄 | 1,213 |
| なに | 問題、仕事、言葉、意味、気持ち、内容 | 12,725 |
| どれだけ | ほとんど、半分、すべて、多少、半数 | 21 |

### 4.2 疑問詞タグの重複付与

予め予想されていたように、複数のタグが重複されて付与された単語があった。この様子を表2に示す。表2から、主に「なに」タグが「いつ」「どこ」「だれ」と重複していることが分かる。

表2:2つの疑問詞が重複した語数

| 疑問詞 | 例 | 件数 |
|---|---|---|
| いつ＋どこ | 現代、飛鳥 | 2 |
| いつ＋なに | 夜、夏休み、クリスマス | 31 |
| どこ＋だれ | 警察、警官、アルバイト、バイト | 4 |
| どこ＋なに | 国、学校、会社、部屋 | 1,350 |
| だれ＋なに | 先生、担当、サラリーマン、弟子 | 15 |

### 4.2 疑問詞タグの付与されない語

前項とは反対に、どの疑問詞タグも付与されない語も 566 語あった。これは「ため」「方」「事」などといったいわゆる形式名詞の他、表3に示すような準形式名詞とでも呼べるような機能性の高い名詞が多く見られた[4]。

これらの名詞は普通名詞でありながら照応の先行詞や質問応答の回答になることが決してなく、従って他の名詞とは明確に異なる扱いを行う必要がある。これらの語の多くは概して高頻度であることから推測すると、様々なタスクで表3のような語を表1のような語と議論しない。

[4] これらとは別に、低頻度語を中心に意味不明の名詞も多数あるが本質的ではないのでここでは

同様に処理を行うことで様々な悪影響を与えている可能性がある。これらの語群は品詞情報からもシソーラス上の分類からも特定できない。従って、例えば本研究で行ったような内省作業が必要であり、これによって名詞の処理対象から排除すべきである。

表3:疑問詞タグの付与されない語の例

| |
|---|
| 場合、ため、方、他、一緒、程度、はず、絶対、普通、利用、一般、互い、一切、OK、方面 |

## ５．　今後の課題

本研究では主要名詞に疑問詞タグの付与を行った。この作業は1名で集中的に作業した結果なので作業結果には若干の誤りが予想され、また個人差も本質的に存在するはずである。これら辞書の品質に関しては継続して改善していく。また 4.1 節で述べたように被覆率(付与名詞数)の改善も行ってゆく。

なお、我々はこの情報で意味的な情報付与が十分だとは全く認識していない。特に「何」タグが付与された名詞群についてはさらに細かく情報付与を行う必要があると考えている。ただし、従来のシソーラスのように、汎用目的という名のもとに様々な意味情報をむやみやたらに付与するつもりはない。様々なタスクへの必要性を吟味した上で真に必要な意味情報は何かを考え、必要に応じて付与を進めていく。

本研究では普通名詞に対して基本的な特性情報を付与したが、我々は以前に動詞に対しても基本的と言える特性情報(意味類型)を付与した(岡田ら 2015;p.497)。雪だるまプロジェクトでは今後も基本的な言語情報を整備し、より高度な自然言語処理の実現を目指す。

## 使用したツールと言語資源

- 日本語解析システム「雪だるま」, (Yamamoto et al. 2015), 長岡技術科学大学 自然言語処理研究室, http://snowman.jnlp.org/
- 現代日本語書き言葉均衡コーパス(BCCWJ), Ver.1.1, 国立国語研究所.


## 謝辞

本研究は、平成 27～31 年科学研究費補助金基盤(B) 課題番号 15H03216、課題名「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」の助成を受けています。


## 参考文献


- 岡田 正平, 山本 和英. 評判分析における品詞情報と意味類型情報の有効性比較. 言語処理学会第 21 回年次大会, pp.497-500 (2015.3)
- 叶内 晨, 小町 守, 岡崎直観, 荒牧英治, 石川博. 風邪に罹ったのは誰か？ － 疾患・症状を保有する主体の推定. 言語処理学会第 21 回年次大会, pp.206-209 (2015.3)
- 柴木 優美, 永田 昌明, 山本 和英. カテゴリ名と記事名の意味属性分類に基づく Wikipedia からの上位下位関係オントロジーの構築. 自然言語処理, Vol.19, No.4, pp.229-279, 言語処理学会 (2012.12)
- 竹野 峻輔, 松田 真希子, 梶原 智之, 山本 和英. 機械学習を用いたニ格深層格の自動付与の検討. 言語処理学会第 20 回年次大会, pp.1011-1014 (2014.3)
- 松井 兵庫, 阪本 浩太郎, 松永 詠介, 神 貴久, 渋木 英潔, 石下 円香, 森 辰則, 神門 典子. 大学入試の穴埋め問題を解く質問応答システムの検討. 言語処理学会第 21 回年次大会, pp.175-178 (2015.3)
- 山本 和英, 宮西 由貴, 高橋 寛治, 猪俣 慶樹, 須戸 悠太, 三上 侑城. 日本語解析システム「雪だるま」 ～ 単語解析部の設計思想 ～. 電子情報通信学会 テキストマイニングシンポジウム, 信学技報, Vol.115, No.222, pp.13-18 (2015.9)
- Kazuhide Yamamoto, Yuki Miyanishi, Kanji Takahashi, Yoshiki Inomata, Yuki Mikami and Yuta Sudo. What We Need is Word, Not Morpheme; Constructing Word Analyzer for Japanese. Proceedings of the International Conference on Asian Language Processing (IALP 2015), pp.49-52 (2015.10)
- Kazuhide Yamamoto and Kanji Takahashi. Construction of Japanese Semantically Compatible Words Resource. Proceedings of the International Conference on Asian Language Processing (IALP 2015), pp.61-64 (2015.10)